Universidad Autónoma Metropolitana
Casa abierta al tiempo Azcapotzalco

Revista Tendencias en Docencia e Investigación en Química 2017

Año 3
Número 3

Percepción de estudiantes de Biología de la FES Iztacala con respecto a sus calificaciones: confiabilidad y validez del instrumento

Chirino Galindo Gladys[1], Gaona Uribe Jessica Guadalupe[1], Urbieta Ubilla Beatriz Rosalia[2], Palomar Morales Martín[1]*

[1]Universidad Nacional Autónoma de México. Laboratorio de Metabolismo de la Diabetes Mellitus, Unidad de Morfología y Función. Facultad de Estudios Superiores Iztacala. Avenida de los Barrios No. 1. Los Reyes Iztacala, Tlalnepantla, Estado de México, C. P. 54090, Estado de México.
[2]Universidad Nacional Autónoma de México, Área de Laboratorio de Investigación Científica II, Carrera de Biología, Facultad de Estudios Superiores Iztacala, Avenida de los Barrios Número 1, Los Reyes Iztacala, Tlalnepantla, Estado de México, C. P. 54090. Estado de México.

*Autor para correspondencia: martin_palomar2004@yahoo.com.mx

Recibido:
17/junio/2017

Aceptado:
14/agosto/2017

Palabras clave
Validez, percepción, calificaciones

Keywords
Validity, perception, qualification

RESUMEN

Los alumnos universitarios frecuentemente tienen la percepción de que es más importante la calificación que el aprendizaje, por varias razones. Para tratar de explorar la percepción de los estudiantes de biología de la FESI con respecto a sus calificaciones, se diseñó un instrumento, que se validó por contenido, criterio, aplicación y constructo. Se obtuvo un instrumento con 26 reactivos agrupados en cinco factores, con una confiabilidad de 0.864, una concordancia Kappa de Fleiss de 0.72, un KMO de 0.885 y una varianza explicada de 59.23%. Las propiedades del instrumento para medir la percepción son adecuadas; en cuanto a los alumnos, la mayoría perciben como importantes las calificaciones, así como un método para promover el aprendizaje y las competencias. Además, a la mayoría de los alumnos les importa más el aprendizaje, la creatividad y el sentido crítico, y desean obtener altas calificaciones sin comprometer sus valores éticos.

ABSTRACT

University students often have the perception that school grades are more important that learning, this due to several reasons. In order to explore the perception about school grades in students of biology of the FESI, an instrument was designed, that was validated for content, criteria, applicability and construct. An instrument was obtained, with 26 items, that were grouped in five factors, with a reliability of 0.864, a Fleiss´ Kappa concordance of 0.72, a KMO of 0.885 and a explained variance of 59.23%. The properties of the instrument for the measuring of perception are accurate; and with respect to the students, most of them perceive school grades as important, as well as a method for promote learning and competences. Also, most of the students are more interested in learning, creativity and critical sense, and they want to get high grades without compromising their ethical values.

Introducción

En general, los estudiantes ingresan a la Universidad con grandes expectativas de movilidad social, pero una vez en el nivel universitario, no buscan la profundidad en sus niveles de conciencia a través del conocimiento (Castrejón, 1990); en lugar de esto, compiten por calificaciones, por plazas en cursos o por admisión en programas de intercambio (Harackiewicz et al., 1998), de manera que un mejor rendimiento estudiantil se asocia principalmente a las calificaciones, las que representan el indicador más empleado (Anaya, 1999; Mathiasen, 1984; Tejedor y García-Valcárcel, 2007).

Cuando se emite una calificación, ésta constituye un multiplicador y reforzador del éxito o fracaso escolar (López, 2005). La asignación de calificaciones en la escuela es una herencia del siglo XIX a la pedagogía, que no está ligada al aprendizaje; por el contrario, está más cercana al poder y al control (Díaz, 1994). Calificar consiste en asignar un número a las actividades del proceso educativo, lo que se traduce en una expresión cuantitativa de resultados de una medición o apreciación; a menudo esta asignación de calificaciones no cuenta con criterios de calidad académica que sean claros, pertinentes y rigurosos (Aguilar et al., 2009); no sólo se trata de reducir la evaluación a la sumatoria de notas y calificaciones que dejan poco margen a la reconstrucción de conocimientos, a la autoevaluación y al fomento de la autonomía por parte del alumnado. Por lo mismo, las calificaciones como medida de rendimiento y evaluación han sido objeto de diversas críticas. En primer lugar, las calificaciones no implican necesariamente calidad ni profundidad de los conocimientos adquiridos, pues tanto un aprendizaje significativo como uno memorístico pueden conducir a calificaciones elevadas (Rodríguez y Ruiz, 2011).

Por otro lado, se define a la percepción como el proceso cognitivo de la conciencia que consiste en el reconocimiento, interpretación y significación para la elaboración de juicios y opiniones en torno a las sensaciones obtenidas del ambiente físico y social, en el que intervienen otros procesos psíquicos entre los que se encuentran el aprendizaje, la memoria y la simbolización (Vargas, 1994).

Todo este acervo incide sobre la forma de entender y llevar a la práctica la actividad de evaluar con la de calificar (Gil-Flores y Padilla, 2009). Además, determinó al profesor y al alumno a la ejecución de juicios de valor; por lo que conceptualizar la percepción sobre calificaciones no es fácil, precisamente porque está implicada a la emisión de dichos juicios (Alegre, 2006). Tampoco podemos dar por hecho que al calificar, confluyen multitud de sensaciones y valoraciones de los estudiantes sobre las que no siempre se ha puesto la suficiente atención (Ricoy y Fernández-Rodríguez, 2013). Por ello, pensamos que el acercamiento a la realidad de los participantes, nos permite un acercamiento con sus pensamientos y juicios que resulta de utilidad para tener conciencia de la problemática que es el poner una calificación.

El interés por la percepción que tienen los universitarios con respecto a una calificación no es algo nuevo, pero no hay estudios sobre percepción alrededor de este tema. De hecho, hay pocos trabajos realizados con el fin de conocer la percepción de los estudiantes universitarios con respecto a su evaluación; mientras que los trabajos realizados con respecto a las calificaciones son mucho más escasos. Trillo y Porto (1999) estudiaron la percepción de los estudiantes sobre su evaluación y calificaciones, y concluyeron que ésta cumple con muy pocas características de validación educativa, y el 95% de la muestra piensa que los padres y la sociedad en general valoran más los diplomas y las calificaciones que el aprendizaje y el esfuerzo. Rodríguez y Ruiz (2011) estudiaron dos indicadores de rendimiento de estudiantes universitarios: calificaciones vs créditos acumulados, y encontraron que el empleo conjunto es una mejor manera de indicadores de rendimiento, pero que uno no sustituye al otro. Silva et al. (2013), estudiaron la percepción de éxito escolar en estudiantes universitarios, y demostraron que las calificaciones contribuyen a ello.

Con el fin de saber cuál es la percepción de los universitarios con respecto a sus calificaciones, y poder dar así cabida a futuros estudios que profundicen más en éstos problemas, se diseñó y aplicó un instrumento. El propósito de la presente investigación es analizar el nivel de validez y confiabilidad de dicho instrumento, así como recabar datos de la percepción de calificaciones en estudiantes de la Carrera de Biología de la Fes Iztacala, mediante su aplicación.

Metodología

La metodología fue tanto cuantitativa (puesto que se recurrió a métodos y estrategias estadísticas con la finalidad de correlacionar y analizar variables para dar respuesta a la validez y confiabilidad del instrumento); como cualitativa (ya que se recurrió a estrategias que valoraban atributos del alumnado); en cuanto a su temporalidad, ésta fue de carácter transversal o transaccional, ello porque se recolectaron los datos en un solo momento con el propósito de conocer su funcionamiento en el instante de la aplicación del instrumento.

Construcción del instrumento

Se establecieron los objetivos a conseguir por cada reactivo, definiendo a este como una pregunta a contestar, afirmación a valorar, problema a resolver, característica a cubrir o acción a realizar. Estos reactivos deben estar siempre contenidos en un instrumento de evaluación específico; y tienen la intención de provocar o identificar la manifestación de algún comportamiento, respuesta o cualidad (Linn y Gronlund, 2000). Los reactivos se organizaron en un cuestionario, y por último se procedió a la definición de los reactivos. Para esta fase se siguieron las indicaciones de Alaminos y Castejón (2006): Cada reactivo debe plantear un solo tema, los reactivos deben ser claros, simples y concisos, evitar reactivos duplicados o no lo suficientemente excluyentes, las palabras utilizadas deben tener el mismo significado para todos los entrevistados, el vocabulario debe ser adecuado y accesible a la población objeto de estudio, y deben evitarse los reactivos tendenciosos. A los reactivos que se originaron se les agrupo en tres categorías. Se optó por un tipo de reactivos redactados en forma de frases o proposiciones, que bien en positivo o en negativo, expresaran ideas acerca del objeto de estudio, que debían ser valoradas de acuerdo a una escala Likert de 5 grados: 1=Totalmente en desacuerdo; 2=En desacuerdo; 3=Ni de acuerdo ni en desacuerdo; 4=De acuerdo y 5=Totalmente de acuerdo; para los reactivos de las dos primeras categorías, un ejemplo de la primera categoría es la afirmación "En determinados contextos, deben preferirse las evaluaciones personales de índole cualitativo no asociadas a calificaciones numéricas". Para la última, los grados de la escala fueron: 1=Nunca; 2=Algunas veces; 3=Frecuentemente; 4=Siempre; un ejemplo de esta categoría es "La calificación es el resultado de tu esfuerzo". Se asignó a cada reactivo una puntuación de uno a cinco o de uno a cuatro según las dimensiones y; viceversa según correspondan, en torno a su dirección, quedando conformado un instrumento provisional de 32 reactivos.

Validez de contenido

La validez de contenido se realizó mediante el criterio de diez expertos con grado de Doctorado en Pedagogía y más de 10 años en al ámbito de la educación, ya que Hyrkäs et al. (2003) manifiestan que este número de expertos brindan una estimación confiable de la validez de contenido de un instrumento. Los expertos recibieron el cuestionario; así como la guía para su validación. Ésta guía se estructuró en dos partes: la primera parte, estaba compuesta por una carta de presentación, las instrucciones para el proceso de respuesta, y las preguntas. La segunda parte permitía apreciar la valoración sobre la validez de contenido del instrumento realizada por cada juez. Concretamente se le solicitó a los expertos que valorasen en una escala del 1 al 5 (1=Nada; 2=Poco; 3=A veces; 4=Bastante; 5=Mucho) de acuerdo a Moriyama (1968); con el objetivo de discriminar si existía diferenciación clara de las categorías a través de las afirmaciones, si había suficiencia y relevancia de reactivos y si eran coherentes entre ellos, también si los enunciados en su conjunto medían la percepción con respecto a las calificaciones y si fueron elaborados con un lenguaje claro, preciso y comprensible para el nivel de los estudiantes. Además, al final de cada uno de los apartados, había en un espacio para las observaciones oportunas. El análisis resultante permitió seleccionar de los 32 reactivos, los mejores en cuanto a su homogeneidad. De aquí se eliminaron 6 y se conformó un instrumento definitivo que constó de 26 reactivos, y se procedió a la aplicación.

Validación por estudio piloto

La aplicación del instrumento se realizó en la FES Iztacala en el mes de abril de 2017. Se tomó una muestra no aleatoria de 100 alumnos voluntarios, con un rango de edad de 18 a 24 años, de segundo semestre de la Carrera de Biología. Se les explicaron los objetivos del proyecto, y se les solicitó responder el cuestionario de manera responsable El tiempo de aplicación fue entre 20 y 25 minutos. Los resultados se sometieron a análisis estadístico básico.

Confiabilidad

Para el cálculo de la confiabilidad del instrumento se empleó el análisis de consistencia interna en las respuestas correspondientes a cada uno de los reactivos que lo conforman, determinando dicha consistencia a través del Alfa de Cronbach.

Validez de criterio

Ante la ausencia de una prueba que reflejara las categorías consideradas por este grupo, se tomó el criterio emitido por 5 profesores del Claustro de Laboratorio de Investigación II, en calidad de informadores claves (Pineault y Daveluy, 1989) que reunieran como requisito la permanencia estable de al menos 5 años en su puesto de trabajo, lo que garantiza un conocimiento profundo de los estudiantes y su percepción con respecto a las calificaciones. Para este aspecto, se utilizó la prueba Kappa de concordancia de Fleiss (Fleiss, 1981), el cual generalizó la aplicación del índice Kappa de Cohen para medir el acuerdo entre más de dos codificadores o informadores para datos de escala nominal y ordinal, dado que este estudio tenía 5 codificadores. Por estas

razones se empleó el indicador Kappa de Fleiss, ya que éste parte de la misma fórmula que propone Cohen, pero generalizada para más de dos codificadores. El coeficiente Kappa de Fleiss añade el cálculo del sesgo del codificador (precisión-error) y el cálculo de la concordancia (calibración) (Torres y Perera, 2009). Se consideró importante si los valores del indicado K eran iguales o mayores a 0.60, lo cual representa una buena o muy buena fuerza de concordancia entre expertos (Blanco-Sánchez, 2014).

Validez de constructo

Un constructo es una conceptualización que requiere de un marco teórico para ser definido. En general, las conceptualizaciones que estudian los investigadores científicos presentan la característica de que no existe un claro consenso a nivel social en cuanto a cómo definirlos o medirlos, sino que para lograrlo se debe contar con una teoría que los sustente. Otra característica fundamental de los constructos es que no son directamente observables y su "captura" a nivel empírico requiere generalmente de rigurosos procedimientos. Son constructos la inteligencia, la personalidad, la creatividad, la percepción, etc., (Babbie, 2000; Kerlinger, 1988). La validez de constructo, se realizó a través del análisis de Componentes Principales (ACP), método que se incluye dentro de los empleados en la llamada validez factorial (Montero, 2008). Con este análisis se pretende someter a prueba los constructos teóricos o dimensiones que generaron el instrumento; se utiliza para ello una técnica de reducción de la dimensionalidad; que posibilita mediante la elección de ejes o componentes "óptimos" reproducir en la medida de lo posible la estructura de los datos originales. Para la elección de los componentes a retener se utilizó el criterio de normalización de Káiser (Bernal et al., 2004), el cual consiste en retener los factores cuyos valores propios sean superiores a 1 (>1). Posteriormente se ajustó el modelo mediante la prueba de esfericidad de Bartlett, cuya finalidad es la de establecer si existen correlaciones distintas de cero en la matriz de correlaciones inicial, supuesto necesario para la realización de un análisis factorial y que debe de utilizarse antes de comenzar el proceso del mismo (Alaminos y Castejón, 2006). Con el propósito de optimizar la interpretación de los resultados se realizó la rotación de los componentes, para ello se empleó el método de rotación Varimax, cuyo objetivo es reducir los componentes en el sentido de conseguir que cada factor o componente rotado tenga unas cargas factoriales o correlaciones altas, sólo para unas pocas variables. Se consideró una correlación importante a partir de 0.4 en concordancia con el límite prefijado por otros autores (Vázquez y Sanz, 1997).

Percepción de los universitarios con respecto a las calificaciones

Una vez aplicado y validado el instrumento, se totalizaron los puntajes con las respuestas de los estudiantes por cada reactivo, y para realizar un análisis de discriminación entre las respuestas para medir la percepción con respecto a las calificaciones, se tomó en cuenta la prueba de la mediana (Alaminos y Castejón, 2006), el cual diferencia entre los puntajes altos y bajos de cada uno de los reactivos, y el puntaje global del instrumento, por lo cual se estableció un punto de corte, considerando la mediana de la sumatoria total de las respuestas, y los cuestionarios resueltos se agruparon en "arriba de la mediana" y "abajo de la mediana". Posteriormente, para cada respuesta se agruparon también los datos arriba y abajo de la mediana. Por último, se realizó un análisis de chi-cuadrado por cada respuesta con respecto a las respuestas totales del área, con el fin de determinar, para cada respuesta, que tan alejado se encontraba el grupo con respecto a la calificación global del área.

Análisis estadístico

El registro y almacenamiento de los datos del instrumento se realizó en una hoja de cálculo Microsoft Excel ® 2013. El análisis de los datos se realizó en el paquete estadístico SPSS ®versión 23.0, con un índice α de 0.05 ó de 0.01de significancia, según se indique.

Resultados y discusión

Construcción del instrumento y validez de contenido

Las valoraciones realizadas por los expertos fueron positivas, teniendo un porcentaje de 90% en acuerdos, coincidiendo en su mayoría en la disminución del número de reactivos o la revisión de agrupación de estos según las categorías o dimensiones. El 81.25% de los reactivos obtuvo calificaciones superiores al 75% en los principios de Moriyama, siendo los reactivos mejor evaluados, es decir que obtuvieron un 100% de respuestas en la categoría "mucho", los siguientes reactivos: 2, 4, 5, 8, 10, 13, 17, 19, 20, 21, 23, 26, 28, 29, 32; los que obtuvieron un 75% de respuestas en la misma categoría fueron los reactivos 1, 6, 9, 12, 15, 18, 22, 24, 27, 30, 31; las puntuaciones más bajas las obtuvieron los reactivos 3, 7, 11, 14, 16, 25, con un 60% de respuestas a la categoría "mucho" por parte de los expertos; después de la validación de éstos y partiendo de sus observaciones y aportaciones, se procedió a suprimir y/o modificar diferentes aspectos del instrumento, pero sin que la base de la estructura general se viera alterada. Los cambios derivados de la revisión, dejaron un instrumento

definitivo de 26 reactivos, puesto que se eliminaron los 6 que tuvieron puntuaciones bajas y se volvieron a numerar otra vez. Los reactivos definitivos se clasificaron en tres categorías con nueve afirmaciones para las dos primeras y ocho para la última. Cohen y Manión (1990) dicen que la definición de las afirmaciones en la elaboración del cuestionario es lo más importante en las investigaciones por encuesta. Ésta es una fase difícil y de abundantes revisiones para llegar a considerar la agrupación de cada uno de los reactivos en función a las categorías principales. Según varios autores (Albert, 2007; Alaminos y Castejón, 2006), este tipo de escala supone una codificación de las respuestas en base a un nivel ordinal que obliga al encuestado a situarse en una posición favorable o desfavorable hacia cada reactivo. La validez de contenido indica que los reactivos de un instrumento de medición deben ser relevantes y representativos del constructo para un propósito evaluativo particular (Escobar-Pérez, 2008). Si un 80 % de los expertos han estado de acuerdo con la validez de un reactivo, éste puede ser incorporado al instrumento (Hyrkäs et al., 2003), en este estudio los porcentajes fueron superiores en los acuerdos de los expertos en cuanto al instrumento completo (90%), y con respecto a cada reactivo también, ya que el valor fue de 81.25%.

Validación por estudio piloto

Se aplicó el instrumento a un total de 100 alumnos, lo que representa el 39% de la población total del segundo semestre de la Carrera de Biología; de los cuales 53% fueron mujeres y 47% fueron hombres; la edad fue de 18 a 24 años (la distribución se observa en la Tabla 1). La edad promedio fue de 21 años. El tiempo de aplicación fue entre 20 y 25 minutos. La fase de estudio piloto del cuestionario es considerada como imprescindible en la elaboración de un instrumento, siguiendo las indicaciones de Gaitán y Piñuel (1998), quienes recomiendan que el número de encuestados debe estar entre 30 y 100 para que tenga validez.

Tabla 1. Distribución de edad de los estudiantes del universo muestral.

EDAD	FRECUENCIA
18	19
19	50
20	22
21	11
22	6
23	8
24	4

Confiabilidad

Se empleó la prueba estadística Alfa de Cronbach (Quero, 2010), la cual arrojo un valor de correlación de 0.864, ya que presentó una adecuada confiabilidad en el análisis, con un total de 26 reactivos, indicando que es buena la consistencia interna del instrumento en la percepción de los estudiantes con respecto a la calificación. El resultado de nuestro trabajo es parecido al 0.850 que obtienen Beramendi y Zubieta en 2014, el cual califican de moderado a alto. Para algunos autores, la consistencia interna de una escala se considera aceptable cuando se encuentra entre 0.70 y 0.90, otros sugieren que la consistencia interna de un instrumento es adecuada si el coeficiente alcanza valores entre 0.80 y 0.90, más aún cuando se está en los primeros estadios de construcción de la escala propuesta por Campo-Arias y Oviedo (2008), Esta confiabilidad nos indica que el 84.0 de la variabilidad es cierta y el 16.0 restante probablemente producto del error inherente al tipo de medición, el error no sistemático. Al respecto Overall y Marsh (1980) destacan que la fiabilidad en las encuestas de estudiantes se entiende como el acuerdo relativo (unanimidad) entre valoraciones de diferentes estudiantes bajo la premisa de que cualquier varianza específica del estudiante individual es aleatoria y debería ser considerada como varianza de error.

En la Tabla 2, se observan la estadística descriptiva del instrumento, las correlaciones de los reactivos son significativas (>0.4), entre los 26 reactivos y el puntaje total del instrumento (Arechabala y Miranda, 2002); Aunado a esto, el instrumento se conserva estable entre 0.81 y 0.86, al ir eliminando uno a uno los reactivos.

Tabla 2. Estadísticos para cada reactivo, media (M), desviación estándar (DE), alfa de Cronbach sin el reactivo (ACSR), y correlación reactivo-instrumento (CRI).

Reactivo	M	DE	ACSR	CRI
1	3.34	0.98	0.85	0.603
2	3.4	1.10	0.84	0.586
3	3.28	1.05	0.83	0.402
4	4.04	0.90	0.81	0.451
5	2.6	1.31	0.82	0.436
6	3.62	1.31	0.86	0.516
7	2.68	1.19	0.85	0.694
8	3.4	1.35	0.86	0.554
9	3.88	0.96	0.82	0.489
10	2.40	1.19	0.86	0.654
Reactivo	M	DE	ACSR	CRI

(Continuación)				
11	3.14	1.28	0.85	0.637
12	3.12	1.36	0.82	0.493
13	3.02	1.41	0.81	0.501
14	3.04	1.44	0.84	0.544
15	3.24	1.37	0.82	0.498
16	2.98	1.25	0.85	0.598
17	3.52	1.32	0.81	0.460
18	3.52	1.40	0.85	0.529
19	2.11	0.89	0.86	0.617
20	2.56	0.84	0.84	0.641
21	2.34	0.91	0.81	0.440
22	2.12	0.80	0.82	0.415
23	2.36	0.71	0.82	0.437
24	2.18	0.66	0.83	0.490
25	2.16	0.64	0.82	0.462
26	2.44	0.5	0.85	0.584

Validez de criterio

Este parámetro del instrumento se probó calculando la correlación del criterio externo de los 5 codificadores que contribuyeron a establecerlos, con respecto a las respuestas de los estudiantes. Para este fin se realizó el análisis de Kappa de Fleiss, y se obtuvo un resultado de $k \geq 0.72$ para los 26 reactivos, lo que se cataloga como una fuerza de concordancia considerable o buena, según Landis y Koch (1977) y Torres y Perera (2009), quienes indican que el rango debe estar entre 0.61 a 0.80, y de acuerdo a la interpretación de Fleiss, la fuerza de concordancia es buena si se encuentra en entre 0.61 a 0.75. El resultado en el estudio está en el rango dado, lo que quiere decir que el instrumento se relaciona con un criterio externo, que es una medida fiable del rendimiento que se quiere pronosticar con el instrumento (Abad et al. 2004).

Validez de constructo

Antes de llevar a cabo el análisis factorial, se establecieron los siguientes criterios en la elección de los factores y los reactivos que lo conformarían: El reactivo debía tener una carga factorial igual o superior a 0.50. Se debía incluir en un solo factor; aquel en el que presentara un mayor nivel de carga factorial, de modo que reactivos con cargas similares en distintos factores serían excluidos. Debía tener una congruencia conceptual entre todos los reactivos incluidos en un factor. El componente o factor debía estar conformado por tres o más reactivos, a excepción de aquellos factores en el que dos reactivos estuvieran sustentados por la teoría o por el coeficiente de consistencia interna. Un factor debía poseer una confiabilidad superior a 0.5 (Londoño et al., 2006).

Se sometieron los 26 reactivos que obtuvieron un nivel aceptable de confiabilidad, a un análisis factorial, utilizando el método de los componentes principales con rotación Varimax. Se utilizó el método de componentes principales para identificar el número y composición de factores necesario para resumir las puntuaciones observadas en un conjunto grande de variables observadas (Lloret-Segura et al., 2014). Este método explica el máximo porcentaje de varianza observada en cada reactivo a partir de un número menor de componentes que resuma esa información. Se obtuvo la medida de adecuación de datos para el análisis de factores, llamada KMO por las iniciales de los autores que lo propusieron (Kaiser, Meyer y Olkin) de 0.885 y la prueba de esfericidad de Bartlett llega a 4260.115 ($p<.001$). Según Nava et al. (2015), el KMO fue bueno, lo que fue congruente con usar el análisis factorial como apropiado para esta muestra de datos, después de varias interacciones, a través de la rotación Varimax se generaron 5 factores y una varianza explicada total de 59.23%; de acuerdo con la proporción de varianza que es explicada por los factores comunes, usualmente, al no existir valores próximos a cero, como proponen Latorre y Pantoja (2012), se confirma que los 26 reactivos son explicados por los componentes.

El primero de los factores encontrados, se recabó sobre la valoración ética que el alumno le da a su calificación, integrado por 8 reactivos, que aportan el 29.35% de la varianza total explicada y una confiabilidad de alfa de Cronbach de 0.925. El segundo factor, concerniente a los objetivos de las calificaciones, quedó integrado por 5 reactivos con una aportación a la varianza total explicada de 9.39% y un nivel de confiabilidad de alfa de Cronbach de 0.852. El tercer factor aborda la importancia social que se le da a las calificaciones, aquí se agruparon 4 reactivos, presentando una varianza de 8.31% y un nivel de confiabilidad de alfa de Cronbach de 0.878. El cuarto factor, referido a las reacciones cognitivas que produce el estrés ante las calificaciones, se agrupó con 5 reactivos aportando una varianza de 6.53% y un alfa de Cronbach de 0.846. El quinto factor describe las reacciones físicas que produce el estrés relacionado con las calificaciones, con una varianza de 5.65% y un alfa de Cronbach de 0.821.; como se observa en la tabla 3.

Tabla 3. Matriz de componentes rotados, se aprecian las cargas factoriales y los 5 factores en los que se dividieron.

Componentes	1	2	3	4	5
Reactivo 19	0.791				
Reactivo 26	0.755				
Reactivo 5	0.727				
Reactivo 22	0.796				
Reactivo 25	0.719				
Reactivo 24	0.737				
Reactivo 3	0.765				
Reactivo 20	0.761				
Reactivo 4		0.737			
Reactivo 7		0.689			
Reactivo 9		0.729			
Reactivo 2		0.653			
Reactivo 21		0.677			
Reactivo 6			0.767		
Reactivo 8			0.784		
Reactivo 1			0.728		
Reactivo 23			0.617		
Reactivo 18				0.752	
Reactivo 14				0.737	
Reactivo 17				0.629	
Reactivo 16				0.736	
Reactivo 13				0.547	
Reactivo 11					0.797
Reactivo 15					0.723
Reactivo 12					0.705
Reactivo 10					0.748
Varianza explicada	29.35%	9.39%	8.31%	6.53%	5.65%
Varianza total 59.23%					
Alfa de Cronbach	0.925	0.852	0.878	0.846	0.821

Percepción de las calificaciones

Con respecto a las respuestas, para el factor 2, concerniente a los objetivos de las calificaciones; y el factor 3, la importancia social de las calificaciones; el 83% de los alumnos pertenecían al grupo arriba de la mediana, y el 17% restante lo contestó por debajo de la mediana. Para el factor 4, reacciones cognitivas que produce el estrés; y factor 5, reacciones físicas que produce el estrés, el porcentaje fue de 51 y 49%, respectivamente. Y para la valoración ética, la distribución fue de 96% por arriba y 16% por abajo. Cada pregunta se comparó contra su correspondiente grupo, de acuerdo al factor que pertenecía, mediante la prueba de chi-cuadrada, y se observó, que para los factores 2 y 3, cinco preguntas fueron respondidas de manera individual por abajo de la mediana, y cuatro de manera muy cercana (Figura 1); para los factores 3 y 4, seis preguntas fueron respondidas de manera individual por abajo de la mediana, y tres de manera muy cercana (Figura 2); por último, para el factor 1, sólo tres fueron contestadas de manera individual por debajo de la mediana, y las otras cinco muy cercana a la mediana (Figura 3). Debido a la distribución asimétrica del grupo, en el factor 1, aquellas preguntas en donde el grupo se encontraba por arriba de la mediana (preguntas 22, 23 y 24), fueron consideradas, por el programa estadístico como no diferentes de ésta.

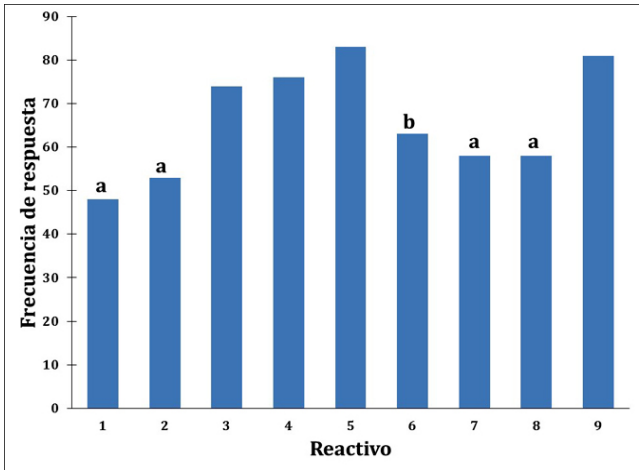

Figura 1. Frecuencia de respuesta, de los factores 2 y 3, objetivos e importancia social de las calificaciones (9 preguntas en total), del universo muestral. Por claridad, sólo se representa la frecuencia de alumnos que contestaron por arriba de la mediana. [a]Diferencia ($p < 0.001$, evaluada por chi-cuadrada) entre la respuesta a la pregunta, y la respuesta global a las preguntas de los factores (83 arriba/17 abajo). [b]Diferencia ($p < 0.01$) evaluada por chi-cuadrada) entre la respuesta a la pregunta y la respuesta global a las preguntas de los factores.

En relación con los resultados más relevantes de respuestas a percepción, el factor 2 registró a la pregunta 1, el 55% de los alumnos perciben que las calificaciones son importantes, porque provocan expectativas en sus padres y en el sistema; a la pregunta 6, el 60% opina que las calificaciones reflejan el modo en que la sociedad comunica sus valores referidos a la educación y las habilidades de los alumnos; con respecto a esto, Trillo y Porto (1999), reportan que el 95% de sus encuestados perciben que sus padres y la sociedad en general valoran más los diplomas y las calificaciones que el aprendizaje y el esfuerzo en nuestro estudio, no se llega al 95% pero

más de la mitad de los estudiantes percibe lo mismo. En relación a la pregunta 8, las calificaciones bajas desmoralizan a los estudiantes, el 57% cree que así es, la motivación en la escuela es un proceso por el cual se inicia y dirige una conducta hacia el logro de una meta.

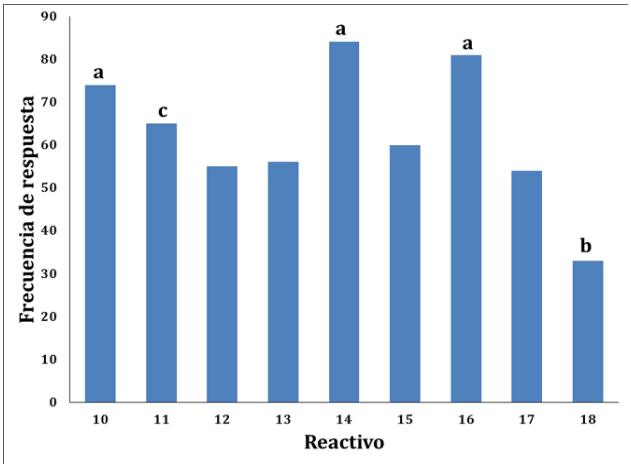

Figura 2. Frecuencia de respuesta, de los factores 4 y 5, reacciones cognitivas y físicas que produce el estrés ante las calificaciones (9 preguntas en total), del universo muestral. Por claridad, sólo se representa la frecuencia de alumnos que contestaron por arriba de la mediana. [a]Diferencia ($p < 0.001$, evaluada por chi-cuadrada) entre la respuesta a la pregunta, y la respuesta global del área. [b]Diferencia ($p < 0.01$) evaluada por chi-cuadrada) entre la respuesta a la pregunta, y la respuesta global al área (51 arriba/49 abajo). [c]Diferencia ($p < 0.05$) evaluada por chi-cuadrada) entre la respuesta a la pregunta, y la respuesta global al área.

Este proceso involucra variables cognitivas y afectivas (Alcalay y Antonijevic, 1987), ambas variables operan juntas para complementarse y hacer eficiente el proceso de aprendizaje (Edel, 2003). En cuanto al factor 3, la pregunta 2 registró que el 51% de ellos piensa que el propósito de las calificaciones es promover el aprendizaje de sus progresos y competencias evaluadas por profesores; a la pregunta 7, el 53% considera que las calificaciones deben basarse en criterios claros y específicos, que midan los logros con respecto a objetivos. En este aspecto el uso de criterios claros y que impulsen los progresos de los estudiantes en una evaluación formativa, que valore la efectividad del proceso de enseñanza-aprendizaje, es fundamental para identificar tanto a estudiantes como profesores sus fortalezas y debilidades, para corregir fallas y lograr los objetivos de aprendizaje (Díaz y Hernández, 2002).

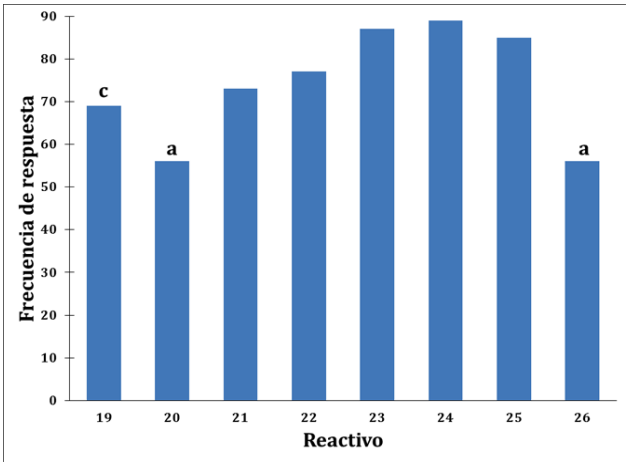

Figura 3. Frecuencia de respuesta del factor 1, valoración ética (8 preguntas) del universo muestral. Por claridad, sólo se representa la frecuencia de alumnos que contestaron por arriba de la mediana. [a]Diferencia ($p < 0.001$, evaluada por chi-cuadrada) entre la respuesta a la pregunta, y la respuesta global a las preguntas del factor (84 arriba/16 abajo). [c]Diferencia ($p < 0.05$) evaluada por chi-cuadrada) entre la respuesta a la pregunta, y la respuesta global al factor.

En relación al factor 5, la pregunta 10 tuvo un 56% de alumnos que se sienten tristes por no obtener una buena calificación; a la pregunta 11, el 68% tiene calificaciones bajas por no realizar tareas; en cuanto al factor 4, la pregunta 14 tuvo un 53% de alumnos que pierden el apetito por preocuparse por calificaciones; a la pregunta 16, el 54% sufre ansiedad por no conseguir buenas calificaciones; y a la pregunta 18, el 85% le preocupa mantener un promedio alto académicamente. El estrés es una condición natural que experimenta el ser humano cuando está excesivamente ocupado o cuando se encuentra bajo presión, como es el caso de los estudiantes en eventos típicos del proceso de enseñanza-aprendizaje-evaluación, y en donde presentan diferentes tipos de estrés y efectos de este, las respuestas al estrés se pueden manifestar tanto en forma fisiológica a través de enfermedades y de manera psicológica (Cohen y Williamson, 1991).

Con respecto al factor 1, las preguntas con diferencia estadística significativa fueron: la pregunta 19, al 60% le interesa más su aprendizaje que su calificación; la pregunta 20, al 51% le importa más la creatividad, reflexión y sentido crítico que un aprendizaje mecánico memorístico; y la pregunta 26, al 59% le interesa obtener buenas calificaciones, pero no comprometiendo sus

valores (Figura 4). Se dice que la evaluación (obviamente, al final del curso se tiene que calificar) en la universidad, debe fomentar la reflexión, la criticidad y la colaboración, no el aprendizaje mecánicamente memorístico, como lo expresan los alumnos, pero si durante el proceso de enseñanza-aprendizaje-evaluación se da más peso en la calificación final a exámenes y cuestionarios, que realmente no midan conocimientos aplicados y habilidades, se seguirá enviando el poderoso mensaje de que lo importante es la repetición y la copia y no la adquisición de conocimientos sólidos (Lacueva, 1997).

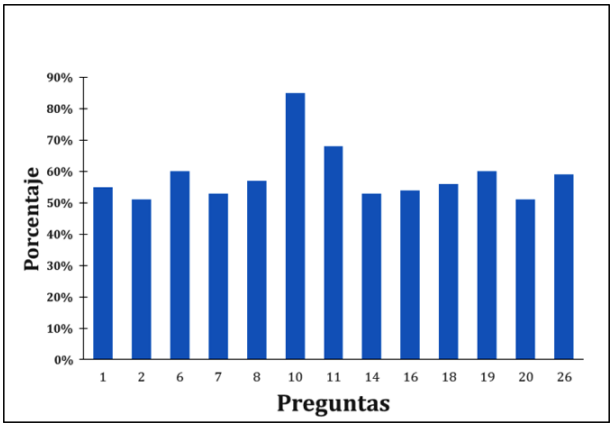

Figura 4. Porcentaje de respuesta de los alumnos a los reactivos del instrumento para medir la percepción a las calificaciones.

En resumen, calificar es una manera de comunicarnos con los alumnos y con la sociedad en concordancia a lo estipulado por el reglamento. Pero hay que hacerlo de la manera adecuada, con todos los elementos que se valoraron en el proceso de enseñanza-aprendizaje-evaluación; así en este transcurso de calificar, no se clasifica a nadie, solo se apoya a los estudiantes a que reconozcan su forma de aprender para continuar su proceso de alcanzar objetivos para su formación profesional. Cuando un alumno consigue una baja calificación, es posible que falte una adecuada planeación didáctica y la ayuda prestada al estudiante haya fallado, no definimos objetivos de aprendizaje adecuados o no lo motivamos lo suficiente. Obviamente hay muchas circunstancias que tienen que ver con el alumno, que no se valoran para calificarlo y escapan a los criterios del profesor (Vega, 2006). Por lo mismo la calificación es con frecuencia un obstáculo, por lo menos así lo perciben los alumnos; pero, indudablemente, mide situaciones transitorias, no aclara fallas y logros, sino sólo categoriza, aunque nadie lo quiera, termina etiquetando, la imagen del estudiante, no se basa en el trabajo concreto realizado y en el grado de comprensión alcanzado, sino que depende totalmente de este juicio ajeno, abstracto y alienante, que es la calificación (Lacueva, 1997).

Finalmente, la calificación expresa tanto el trabajo del profesor, como del estudiante, obviamente que todo tiene que llevar un proceso y evidencias claras de dicho desarrollo individual. Calificar con justicia es difícil, porque se le concede demasiada importancia a las notas obtenidas, como si fueran lo máximo y el fin, y no sólo un herramienta, una manera para estimar si el progreso del estudiante está relacionado con los objetivos del programa.

Conclusiones

Se logró elaborar un instrumento confiable, consistente, estable, con buena fuerza de concordancia, formado por reactivos que tienen un nivel aceptable de confiabilidad cada reactivo tiene propiedades adecuadas. La mayoría de los alumnos percibe como importantes las calificaciones, así como un método para promover el aprendizaje y las competencias. A la mayoría de los alumnos les importa más el aprendizaje, la creatividad y el sentido crítico, y desean obtener buenas calificaciones sin comprometer sus valores éticos.

Agradecimientos

Al Biólogo Guillermo Elías Fernández, por el apoyo con el programa estadístico.

A los profesores del Diplomado de Investigación para el Perfeccionamiento Docente, por el apoyo para la evaluación del instrumento.

A los alumnos y profesores de la Carrera de Biología, de segundo semestre, por su valiosa participación.

Referencias

Abad F.J., García C., Gil B., Olea J., Ponsoda V. Revuelta J. (2004). Introducción a la psicometría. Teoría clásica de los tests y Teoría de la respuesta al ítem. 1ª Ed. Universidad Autónoma de Madrid, p. 75.

Aguilar M.M., Alcántara-Eguren A.R., Morán S.A. L. (2009). La medición del aprendizaje del alumno, a través de la asignación de calificaciones. Un análisis en la universidad iberoamericana Puebla. X Congreso Nacional de Investigación Educativa. Área 1: aprendizaje y desarrollo humanos. 21-25 de Septiembre. Veracruz.

Alaminos C.A., Castejón C.J.L. (2006). Elaboración, análisis e interpretación de encuestas, cuestionarios y escalas de opinión. 1ª Ed. Marfil, pp. 119.

Albert, M.J. (2007). La investigación educativa. Claves teóricas. 1ª Ed. McGraw Hill, pp. 266.

Alcalay L., Antonijevic N. (1987). Variables afectivas. *Rev. Educ.*, 144: 29-32.

Alegre, O. (2006). Evaluación del programa de posgrado «Educar en la Diversidad» por parte de los profesores participantes. *Rev. Educ.*, 340: 299-340.

Anaya G. (1999). College impact on student learning: Comparing the use of self-reported gains, standardized test scores and college grades. *Res. High. Educ.*, 40: 499-526.

Arechabala M.M.C, Miranda C.C. (2002). Validación de una escala de apoyo social percibido en un grupo de adultos mayores adscritos a un programa de hipertensión de la región metropolitana. *Cien. Enferm.,* 8: 49-55.

Babbie E. (2000). Fundamentos de la Investigación Social. 1ª Ed. Thomson, p. 246.

Beramendi M., Zubieta E., (2014). Construcción y validación de la escala de percepción del sistema normativo. *Rev. Mex. Psicol.*, 31: 124-137.

Bernal G.J.J., Martínez M.D.S.M., Sánchez G.J.F. (2004). Modelización de los factores más importantes que caracterizan un sitio en la red. Disponible en: http://www.um.es/asepuma04/resumen/resumen_bernal_martinez_sanchez.pdf.

Blanco-Sánchez J. P. (2014). Validación de una escala para medir la habilidad de cuidado de cuidadores. *Aquichan,* 14: 351-363.

Campo-Arias A., Oviedo H.C. (2008). Propiedades Psicométricas de una Escala: la Consistencia Interna. *Rev. Salud Púb.*, 10: 831-839.

Castrejón D.J. (1990). El concepto de universidad. 2ª Ed. Trillas, p. 209.

Cohen, L., Manion, L. (1990). Métodos de investigación educativa. 2ª Ed. La Muralla, p. 502.

Cohen S., Williamson G. (1991). Stress and infectious disease in humans. *Psychol. Bull.*, 109: 5-24.

Díaz B.A. (1994). Una polémica en relación al examen. *Rev. Iberoam. Educ.,* 5: 161-181.

Díaz B.F., Hernández R.G. (2002). Estrategias docentes para un aprendizaje significativo: una interpretación constructivista. 2ª Ed. McGraw Hill, pp. 459.

Edel N.R. (2003). El rendimiento académico: concepto, investigación y desarrollo. *REICE.,* Disponible en: http://www.redalyc.org/articulo.oa?id=55110208

Escobar-Pérez J. (2008). Validez de contenido y juicio de expertos: una aproximación a su utilización. *Av. Med.,* 6: 27–36.

Fleiss J.L. (1981). Statistical methods for rates and proportions. 1ª Ed. John Wiley and Sons. p. 737.

Gaitán J.A., Piñuel J.L. (1998). Técnicas de investigación en comunicación social. Elaboración y registro de datos. 1ª Ed. Síntesis, p. 311.

Gil-Flores J., Padilla M.T. (2009). La participación del alumnado universitario en la evaluación del aprendizaje. *Educación, XXI*: 43-65.

Harackiewicz J.M., Barron, K.E., Elliot, A.J. (1998). Rethinking achievement goals: When are they adaptive for college students and why? *Educ. Psychol.,* 33: 1-21.

Hyrkäs K., Appelqvist-Schmidlechner K., Oksa L. (2003). Validating an instrument for clinical supervision using an expert panel. *International J. Nurs. Stud.*, 40: 619-625.

Kerlinger F.N. (1988). Investigación del Comportamiento. 4ª Ed. McGraw-Hill, p. 827.

Lacueva A. (1997). La evaluación en la escuela: una ayuda para seguir aprendiendo. *Rev. Fac. Educ.,* 23 (1-2). Disponible en: http://dx.doi.org/10.1590/S0102-25551997000100008

Landis J., Koch G. (1977). The measurement of observer agreement for categorical data. *Biometrics*., 33: 159-174.

Latorre R.P.Á., Pantoja V.A. (2012). Diseño y validación de una escala de percepción del riesgo en actividades físicodeportivas escolares. *Retos,* 21: 25-29.

Linn R.L., Gronlund N. (2000). Measurement and assessment in teaching. 8a Ed. Prentice Hall, p. 574.

Lloret-Segura S., Ferreres-Traver A., Hernández-Baeza A., Tomás-Marco I. (2014). El análisis factorial exploratorio

de los ítems: una guía práctica, revisada y actualizada. *An. Psicol.*, 30: 1151-1169.

Londoño N.H., Henao l.G.C., Puerta I.C., Posada S., Arango D., Aguirre-Acevedo D.C. (2006). Propiedades psicométricas y validación de la Escala de Estrategias de Coping Modificada (EEC-M) en una muestra colombiana. *Univ. Psychol.,* 5: 327-349.

López P.V.M. (2005). La evaluación como sinónimo de calificación. Implicaciones y efectos en la Educación y en la Formación del Profesorado. *Rev. Electr. Interuniver. Form. Prof.*, 8: 1-7.

Montero R.E. (2008). Escalas o índices para la medición de constructos: El dilema del analista de datos. *Av. Med.*, 6: 15–24,

Moriyama I.M. (1968). Indicators of social change. Problems in the measurements of health status. 1ª Ed. Russell Sage Foundation, p. 593.

Nava Q.C.N., Bezies Á.R., Vega V.C.Z. (2015). Adaptación y validación de la escala de percepción de apoyo social de Vaux. *LIBERABIT*, 21: 49-58.

Overall J.U., Marsh H.W. (1980). Students' evaluations of instruction: A longitudinal study of their stability. *J. Educ. Psychol.*, 72: 321-325.

Pineault R., Daveluy C., (1989). La planificación sanitaria: conceptos, métodos, estrategias. 2ª Ed. Masson, p. 64-85.

Quero V.M. (2010). Confiabilidad y coeficiente alpha de Cronbach. *Telos,* 12: 248-252.

Ricoy M.C., Fernández-Rodríguez J. (2013). La percepción que tienen los estudiantes universitarios sobre la evaluación: un estudio de caso. *Educación,* XXI: 321-341.

Rodríguez A.M.N., Ruíz D.M.A. (2011). Indicadores de rendimiento de estudiantes universitarios: calificaciones versus créditos acumulados. *Rev. Educ.*, 355: 467-492.

Silva G.C., Motte N.A., Garcia S.M.F. (2013). Percepción de éxito escolar en estudiantes universitarios. Factores asociados al abandono. Tercera conferencia latinoamericana sobre el abandono en la educación superior. Disponible en: www.alfaguia.org/www-alfa/images/ponencias/clabesIII/LT_1/ponencia_completa_198.pdf

Tejedor F.J., García-Valcárcel M.R.A. (2007). Causas del bajo rendimiento del estudiante universitario (en opinión de los profesores y alumnos). Propuestas de mejora en el marco del EEES. *Rev. Educ.*, 342: 443-473.

Torres G.J.J., Perera R.V.H. (2009). Cálculo de la fiabilidad y concordancia entre codificadores de un sistema de categorías para el estudio del foro online en e-learning. *Rev. Invest. Educ.*, 27: 89-103.

Trillo A.F., Porto C.M. (1999).La percepción de los estudiantes sobre su evaluación en la Universidad. Un estudio en la Facultad de Ciencias de la Educación. *Innov. Educ.,* 9: 55-75.

Vargas M.L.M. (1994). Sobre el concepto de percepción *Alteridades,* 4: 47-53.

Vázquez C., Sanz J. (1997). Fiabilidad y valores normativos de la versión española del Inventario para la Depresión de Beck de 1978. *Clín. Salud* 8: 403-422.

Vega M. (2006). Evaluar y calificar desde el aprendizaje. Una propuesta de innovación con estudiantes universitarios. Disponible en: http://universidadeuropea.es/myfiles/pageposts/jiu/jiu2005/archivos/EVAL/EVAL06.pdf.